Optimizing Dynamic Resource Allocation in Teamwork

Steve W. J. Kozlowski, Ph.D.
Department of Psychology
309 Psychology Building
Michigan State University
East Lansing, MI 48824-1116
E-Mail: stevekoz@msu.edu
Phone: 517-353-8924
FAX: 517-353-4873

and

Richard P. DeShon, Ph.D.
Department of Psychology
306 Psychology Building
Michigan State University
East Lansing, MI 48824-1116
E-Mail: deshon@msu.edu
Phone: 517-353-4624
FAX: 517-353-4873

Final Performance Report

FA9550-07-1-0483

February 2008

Submitted to:

Jun Zhang, Ph.D.
Program Manager, AFOSR/NL
875 N. Randolph Street
Suite 325, Room 4031
Arlington, VA 22203-1768
1-703-696-8421 (office)
1-703-696-8449 (fax)
DSN: 426-8421
jun.zhang@afosr.af.mil

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U. S. Government.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 02/25/2008	**2. REPORT TYPE** Final Progress Report	**3. DATES COVERED** *(From - To)* 06/01/2007 - 11/30/2007

4. TITLE AND SUBTITLE Optimizing Dynamic Resource Allocation in Teamwork	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER FA9550-07-1-0483
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Steve W. J. Kozlowski and Richard P. DeShon	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Michigan State University 301 Administration Bldg East Lansing, MI 48824-1046	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of /NL Scientific Research/ 875 N Randolph St Arlington VA 22203	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFOSR
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release; distribution is unlimited.

AFRL-SR-AR-TR-08-0156

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The proposed research was designed to extend prior AFOSR sponsored research (DeShon, Kozlowski et al., 2004) to model optimal human resource allocation to account for learning, performance, and adaptation for complex and dynamic tasks incorporating individual and team goals. Phase 1 was intended to implement an optimal, multiple-criterion (individual and team goals) reinforcement learning model that would compare human performance to optimal model performance. Phase 2 was intended to extend the model to autonomous decision makers by incorporating "reward" into the decision maker via satiation levels on individual and team goals, with learning and performance compared to the optimal model (Phase 1) and human benchmarks. Phase 3 was intended to extend the model to encompass adaptation to changes in reward structure (development of an adaptive, model-based reinforcement learning approach that would be compared to standard reinforcement learning and human performance). Funding restrictions limited work to 50% of phase 1 effort (6 of 12 months at 50% of original budget). Funding ceased at 6 months. Initial project efforts were devoted to redesigning and redeveloping our individual-team resource allocation simulation to incorporate features necessary to implement the reinforcement learning model and to evaluate potential implementations of the Q-learning algorithm within the simulation. This report summarizes the intended research contribution and progress up to funding termination.

15. SUBJECT TERMS

Resource Allocation; Goal Regulation; Optimal Modeling; Individual and Team Learning, Performance, and Adaptation.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Steve W. J. Kozlowski
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		18	**19b. TELEPHONE NUMBER** *(include area code)* 517-353-8924

Executive Summary

The originally proposed research, spanning 36 months, was designed to extend prior AFOSR sponsored research (DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2004) to develop model foundations for optimal human resource allocation to account for learning, performance, and adaptation in complex and dynamic task environments that incorporate individual and team goal regulation. Phase 1 was intended to implement an optimal, multiple-criterion (individual and team goals) reinforcement learning model for a decision-making simulation that would provide a foundation for comparisons of human performance to optimal model performance. Phase 2 was intended to extend the model to autonomous decision makers by incorporating "reward" into the decision maker via satiation levels on multiple goals or drives (i.e., individual and team goals). Learning and performance in this model implementation would be compared with the optimal model (Phase 1) and human benchmarks. Phase 3 was intended to extend model development to encompass adaptation to changes in reward structure, i.e., development of an adaptive, model-based reinforcement learning approach that would be compared to standard reinforcement learning and human performance.

The proposed research was designed to advance basic knowledge in several areas. From the perspective of human learning and performance, it would have extended current models to multi-criteria optimization and would have provided an approach to rapid adaptation to environmental change (e.g., spontaneous recovery from extinction) that is a major limitation of virtually every existing model of human learning. From a machine learning perspective, it would have provided a mechanism for substantially improving the autonomy of learning agents by incorporating goals or drives with dynamic satiation levels into the learning agent. Thus, the reward value associated with various actions would be a function of both the environment and the current goal state space of the learning agent. Further, our approach to adaptation, via knowledge representation and memory, represented a substantial departure from existing machine learning approaches to the adaptation problem. It was expected that the the proposed research would raise a number of questions that would spur further machine learning research.

Budget restrictions at the time of proposal submission limited work to 50% of the originally proposed phase 1 effort (6 of 12 months at 50% of original budget). Continuation funding for the project was not provided and work ceased with 6 months of work. Initial project efforts were devoted to redesigning and redeveloping our individual-team resource allocation simulation system to incorporate features necessary to implement the reinforcement learning model and to evaluate potential implementations of the Q-learning algorithm within the simulation. This report summarizes the intended research contribution and progress to the point of funding termination.

Optimizing Dynamic Resource Allocation in Teamwork

Steve W. J. Kozlowski and Richard P. DeShon
Michigan State University

FINAL PERFORMANCE REPORT

Problem Background and Research Objectives

Problem Background

One of the important emerging trends in military thought about the future of warfare centers on information superiority: the ability to collect information, make sense of it, and act on it rapidly. Military planners assert that "… the time required by individuals to access or collect the information relevant to a decision or action has been reduced by orders of magnitude, while the volume of information that can be accessed has increased exponentially …. across a broad range of value-creating activities, the fundamental limits to the velocity of operations are no longer governed by time or space. Instead, the fundamental limits are governed by the acts of deciding, by the firing of neurons, by the speed of thought" (Alberts, Garstka, & Stein, 1999, p. 16). Planners speculate that this trend to leverage the power of information technologies through Network Centric Operations (NCO), already well underway, will increase operational tempo and revolutionize command and control operations.

Ensuring the effectiveness of this vision will entail advances in our understanding of humans operating in such systems and, in particular, to the development of theory for modeling and optimizing human performance. NCO task environments place high demands on human learning, performance, and performance adaptation. Such task environments are dynamic, ambiguous, and emergent. They necessitate rapid situation assessment, prioritization of possible actions, and accurate decision making. And, they require performance adaptation as the situation evolves, often with unexpected shifts. Thus, key features of the NCO task environment from the perspective of human functioning include time pressure, technology mediation, the presence of other agents (i.e., NCO involves team networks), and the need to adapt as the task situation changes. *These latter two factors – the need to integrate effort with that of one's teammates and the necessity for adaptation – add considerable complexity to any effort to understand and model optimal human performance in such settings.*

Research Objectives

The purpose of the originally proposed research was designed to develop a model of optimal human resource allocation to account for learning, performance, and adaptation in the task environment described above. Prior research has characterized human performance in such settings as a process of dynamic multiple goal, multilevel resource allocation (DeShon, Kozlowski, Schmidt, Milner, & Weichmann, 2004). The premise of this approach is that self-regulation provides a reasonable account of individual learning and performance. Team tasks, however, necessitate that individuals allocate attention and effort to the pursuit of multiple goals – individual and team. Team members strive to meet their own responsibilities, while also monitoring team performance. Thus, individuals monitor progress on two goal loops – individual and team – and dynamically allocate resources to one goal or the other as needed to maintain overall performance. Research has shown that this process of dynamic resource allocation accounts for both individual and team performance on a complex an interdependent task (DeShon et al., 2004). Thus, it provided a theoretical point of departure to develop an optimal model of dynamic resource allocation.

Three key advances were necessary to achieve this goal. First, it was necessary to focus on mathematically optimal performance as a standard. Second, we needed to incorporate a clear learning process into our modeling of skill acquisition as individuals learn how to allocate limited resources to the achievement of individual and team goals. Third, we needed to function at lower levels of analysis by incorporating control process into our modeling. Recent advances in reinforcement learning provided a vehicle to meet these and additional research goals.

Scientific Approach

We proposed a three year project investigating these research objectives that consisted of three distinct research phases. Phase 1 was intended to implement an optimal multiple-criterion reinforcement learning model that will provide a foundation for comparisons of human performance capability and variability to optimal model performance. The model was intended to be implemented around our experimental platform (TEAMSim), necessitating scenario redesign in parallel with model development. Parameterization of the model was intended to be obtained using both prior research (e.g., Fu & Anderson, 2006) and asymptotic human performance. Phase 2 was intended to extend the reinforcement model by locating multiple, individual and team oriented goals (i.e., reward values) in the decision maker in an effort to more realistically model human resource allocation and learning. This model implementation was intended to be compared with the optimal model (Phase 1) and human performance benchmarks. Phase 3 was then intended to further extend model development to encompass adaptation to changes in the reward structure. Real world environments are typically in flux. Thus, we planned to develop an adaptive, model-based reinforcement learning approach that would have been compared to standard reinforcement learning and human benchmark performance.

Theoretical Foundation

The dynamic planning of actions under conditions of uncertainty is certainly one of the most important problems faced by team members performing complex tasks in stochastic environments. We begin our approach to this problem by first considering the case where there is no uncertainty and then generalize these results to the case where environmental knowledge is virtually nonexistent. When knowledge is complete and accurate the environment can be characterized as a known set of discrete states (***S***), then it is common to model stochastic decision problems as Markov Decision Processes (MDP). Specifically, a MDP is a quadruple (S, $A, P_{ss'}^{a}, r_{s}^{a}$) where S is a finite set of states, A is finite set of actions, $P_{ss'}^{a}$ is a (Markovian) probability of transitioning from state s to s' when action a is undertaken, and r_{s}^{a} is the expected numerical reward that is received immediately upon selecting action a in state s. In this model the next state and the expected reward depend only on the previous state and the action taken. The *discounted, infinite horizon planning task* in a MDP is to determine a decision policy π:

S→*A* that maximizes the value of every state, $V^{\pi}(S) = E\left[\sum_{t=0}^{\infty}\gamma^{t} r_{(t+1)} \middle| s_0 = s, \pi\right]$, where s_0 is the state at time 0, $r_{(t+1)}$ is the reward achieved at time *t*+1, γ is a discount factor in (0,1), and the expectation is taken by following the state dynamics induced by π. The function V_π is called the value function of policy π. The optimal value function *V**, associated with an optimal policy, is the unique solution to the set of equations:

$$V^{*}(S) = \max_{a \in A}\left(r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V^{*}(s') \right) \tag{1}$$

and can be used to determine an optimal policy by choosing actions in a greedy manner. Algorithms that yield a solution to Equation 1, such as value iteration and policy iteration, are widely available and well understood (e.g., Puterman, 1994).

This basic decision model may be generalized in a number of ways to yield more realistic environment-actor representations. If the decision maker is unable to determine the current state with complete reliability, then the Markov Decision Process may be generalized to a Partially Observable Markov Decision Process (POMDP). In this case, the basic MDP model is supplemented with a memory process that represents belief states as probability distributions over state space, *S*. The beliefs are updated as experience is gained through decisions that yield state-action pairings. However, this model still requires that the state transition probabilities, $P_{ss'}^a$, are known. If these probabilities are unknown they must also be learned before optimal policies can be identified. Reinforcement learning is an increasingly common approach that can yield an optimal decision policy even when very little is known about the state-action pairings. Using simple trial-and-error action, reinforcement learning methods enable the discovery of an optimal decision policy.

Reinforcement learning represents a unique intersection of machine, animal, and human learning approaches being investigated in cognitive psychology, differential psychology, artificial intelligence, biology, and optimal control engineering. Reinforcement learning algorithms attempt to find a *policy* that maps the possible actions a decision maker can undertake onto environmental states to maximize one of many long-term reward criteria. If the environmental states and decision maker actions can be formulated as a Markov Decision Process then the action policy identified via reinforcement learning algorithms converges to the optimal decision policy as long as all state-action pairs are thoroughly sampled (e.g., Dayan,1992; Watkins & Dayan, 1992). Reinforcement learning differs from more traditional supervised learning (e.g., Bayes classifier or neural network) in that correct input/output pairs are not presented and sub-optimal actions are not explicitly corrected. Instead, a reinforcement learning decision maker learns from the consequences of its trial-and-error actions and the rewards accrued through the various actions, rather than from being explicitly taught the correct action. Further, there is a focus on real-time performance that requires a balance between exploring unexamined state-action pairs versus exploiting current knowledge to perform at the best level currently possible.

Temporal Difference Reinforcement Learning

Several algorithms exist to implement the reinforcement learning process. The one-step Q-learning algorithm – a variant of temporal difference learning – is one of the most commonly used reinforcement learning algorithms and has been shown to converge to an optimal policy under realistic conditions. Given a particular state, *s*, a learning decision maker receives a specific reward, r_s^a, by undertaking a particular action, *a*. However, it is important to consider both the immediate reward that is received and any future reinforcements that result from ending up in a new state where further actions can be taken that follow a particular policy. The Q-value for a state-action pair is the sum of all of these reinforcements, and the Q-value function is the mapping of state-action pairs to values. The optimal policy, ***Q****, for selecting an action given the current state may be represented as the Q-value function,

$$Q^*(s_t, a_t) = r_{s_t}^{a_t} + \gamma \max_{a_{(t+1)}} Q^*(s_{(t+1)}, a_{(t+1)}) \tag{2}$$

where ***t*** is a time index, and **γ** is a discounting weight that determines the relative importance of current and future rewards. A higher value of **γ** means that the future matters more for the Q-value of a given action in a given state. In words, the optimal Q-value of for a particular action in a particular state is the sum of the reward received when that action is taken and the discounted best Q-value for the state that is reached by taking that action.

If the Q-values of every state-action pair were known, a decision maker could use this information to select an action for each state to maximize reward. The problem is that in an unknown environment the decision maker initially does not know the Q-values of any state-action pairs. Therefore, the decision maker's goal is to learn the state-action Q-values that yield optimal reward. At any given time during learning, the decision maker stores a particular Q-value for each state-action pair. At the beginning of learning, the Q-values are set to either a random variate or a default value. Learning should move the stored Q-values closer to the optimal values ($\boldsymbol{Q}^*$). To do this, the decision maker repeatedly takes actions in particular states and notes the reinforcements that it receives. The stored Q-value is then updated for that state-action pair using the reinforcement received. Assuming the Q-values are stored in a lookup table, the Q-learning update function is:

$$Q^{new}(s_t, a_t) = (1-\alpha)\, Q^{old}(s_t, a_t) + \alpha \left[r_{s_t}^{a_t} + \gamma \max_{a_{(t+1)}} Q^{old}(s_{(t+1)}, a_{(t+1)}) \right] \tag{3}$$

where **α** is a learning rate parameter that limits the learning step size that occurs in the updating process. The new Q-value for the state-action pair is the weighted combination of the old Q-value for that state and action and the newly acquired reward information.

A fundamental aspect of reinforcement learning is that learning occurs through a prediction error process. Current Q-values provide predictions of the best action for a given state. Once the action is taken, a reward is obtained and any difference between the predicted reward for the state-action pair and the actual reward is used to update the Q-value for future predictions. Nothing is learned about a state-action pair unless the action is attempted in the given state so that the expected reward can be compared to the actual reward. Further, a decision maker should attempt a range of actions, including actions predicted to be suboptimal, to determine the most effective action for a specific state. This leads to a trade off between exploiting current knowledge to select the action that is expected to yield the highest reward in a given state or selecting an action at random to explore the possibility that alternative actions might result in higher reward. Exploitation is based on what the decision maker knows about the reward structures in the environment and has a higher probability of reward. On the other hand, exploration makes it possible to discover actions that would not otherwise be tried that could actually result in higher immediate or cumulative reward. In Q-learning, the trade off between selecting actions that reflect either exploration and exploitation is typically determined using either an ε-greedy or a softmax strategy based on a Boltzmann probability distribution (Sutton & Barto, 1998). The softmax criterion used in this research will select an action, ***a***, from a set of potential actions, ***A***, with probability,

$$\frac{e^{Q(a_t)*age/\tau}}{\sum_{b=1}^{n} e^{Q(b_t)*age/\tau}}, \tag{4}$$

where τ is a parameter controlling the exploitation rate, *age* reflects the duration of the learning process, and *b* reflects the actions other than the given action in the set of potential actions. As the ratio of *age*/τ gets larger, actions are selected in an increasingly exploitive manner that improves the rate of convergence. Early in learning, it is important to explore because the system knowledge is unreliable. However, as system knowledge increases, exploration becomes increasingly ineffective. Initial parameterization of the reinforcement learning algorithms will be based on recent work by Fu and Anderson (2006), who provide cognitively reasonable values for the parameters in this model.

In addition to overcoming many limitations associated with standard Markov Decision Processes, both the theory underlying reinforcement learning and the algorithms used to implement reinforcement learning are strongly connected to existing research on actual human learning and motivation processes. The concept of prediction error and the brain's sensitivity to prediction error underlies Rescorla and Wagner's (1972) model of associative learning. In this model, learning occurs not because two events simply covary, but rather because the observed covariance is unanticipated based on the current associative strength between the events. As such, the Rescorla-Wagner model is fundamentally an error-correction model closely connected to temporal difference learning algorithms in reinforcement learning (Sutton & Barto, 1998). Specifically, the Rescorla-Wagner model specifies that

$$\Delta V_X^{t+1} = \alpha_X \beta \left(\lambda - V_{total}^t \right) \tag{5}$$

and

$$V_X^{t+1} = V_X^t + \Delta V_X^{t+1} \tag{6}$$

where ΔV_X^{t+1} is the change in the associative strength (*V*) of the conditioned stimulus, *X*, as a result of pairing with the unconditioned stimulus (US_1) on trial *t*+1, α_X is a rate parameter for the CS and ranges from 0 to 1, β is a rate parameters for the US and also ranges from 0 to 1, λ is the maximum conditioning US_1 can produce and represents the limit of learning, and V_{total}^t is the sum of associative strengths of all CSs that are present on trial t+1. In Equation 6 ΔV_X^{t+1} is the associative strength of CS *X* after trial *t*+1, V_X^t is the associative strength of CS *X* before trial *t*+1.

On any given trial the current associative strength, V^{t}_{total}, is compared with λ and the difference is treated as an error to be corrected by producing a change in associative strength (ΔV). This model accounts for many classic phenomenon observed in human and animal learning such as acquisition rate, extinction, discrimination, overshadowing, and blocking (Miller, Barnet, & Grahame, 1995).

In addition to learning models, recent behavioral neuroscience evidence, initiated by Shultz and colleagues (e.g., Shultz, 1998; Shultz, 2002; Schultz, Dayan and Montague, 1997), has demonstrated a direct bridge between reinforcement learning and midbrain dopamine neurons. In reinforcement learning algorithms, prediction errors are used to learn state-action pairings that optimize future rewards. A surprising similarity is found in the functioning of dopamine neurons in the brain. The ventral tegmentum and the substantia nigra are part of the brain's pleasure system and are thought to be one of the major sources of reward and motivation in the brain. Dopamine neurons in these areas have been shown to be involved when prediction errors are encountered in learning experiments for both animals and humans. This finding has been replicated and extended in many subsequent investigations (e.g., Montague & Berns, 2002; Montague, Hyman, & Cohen, 2004; Smith, Becker, & Kapur, 2006; Wise, 2004).

Intended Research Plan

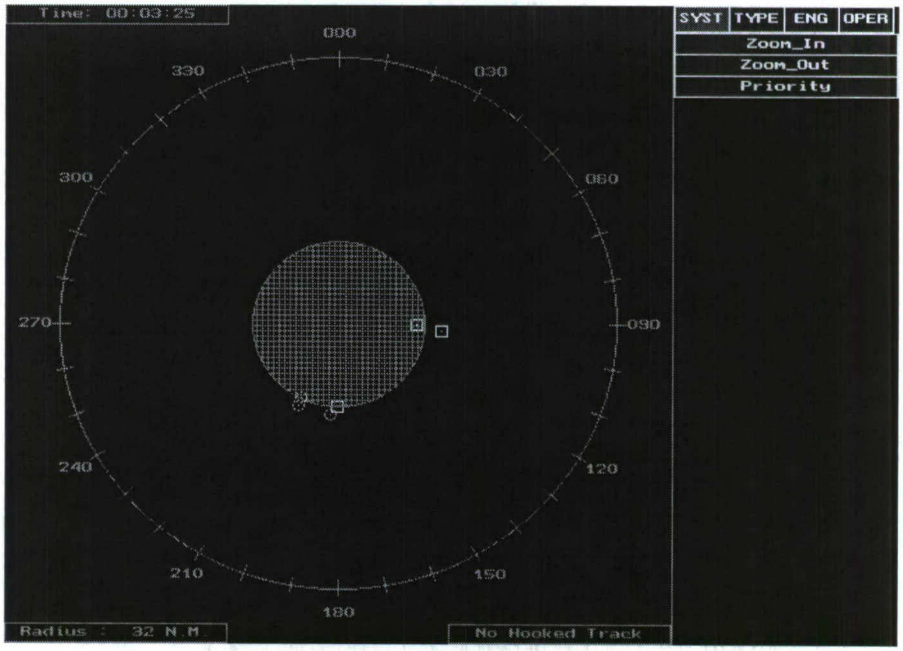
The basic reinforcement learning process described above provides a psychologically relevant optimal standard that may be used to determine the effectiveness of various interventions designed to both push human decision makers toward optimal resource allocation and to reduce the variance human performance. A number of interesting and important complexities arise when using reinforcement learning to model human knowledge acquisition and resource allocation (e.g., origins of reward, non-stationary reward structures, adaptation to environmental reward contingencies). These complexities provide opportunities to contribute to the knowledge of the human resource allocation processes, to advance general theories of human learning, and to potentially provide new structures for knowledge representation in the machine learning literature. Phase 1 of the proposed research was intended to implement an optimal, multiple-criterion (individual and team goals) reinforcement learning model consistent with our experimental platform (TEAMSim) that would provide a foundation for comparisons of human performance capability and variability to optimal model performance. Phase 2 was intended to extend the reinforcement model to autonomous decision makers by incorporating "reward" into the decision maker via satiation levels on multiple goals or drives (i.e., individual and team oriented goals). Although this represents a departure from the dominant machine learning perspective, it is a natural -- and likely necessary -- implementation for modeling human behavior. Learning and performance in this model implementation was intended to be compared with the optimal model (Phase 1) and human benchmarks. Phase 3 was then intended to extend model development to encompass adaptation to changes in the reward structure; we would have developed an adaptive, model-based reinforcement learning approach that would be compared to standard reinforcement learning and human performance.

As noted previously, initial funding was restricted to 6 months (50%) of the proposed phase 1 effort. There was no continuation funding. Initial project effort during the 6 months of support were directed toward modifying our simulation system to incorporate features necessary to implement the reinforcement learning model and to evaluate potential implementations of the Q-learning algorithm within the simulation. These developments are briefly summarized below.

Phase 1 Work Performed

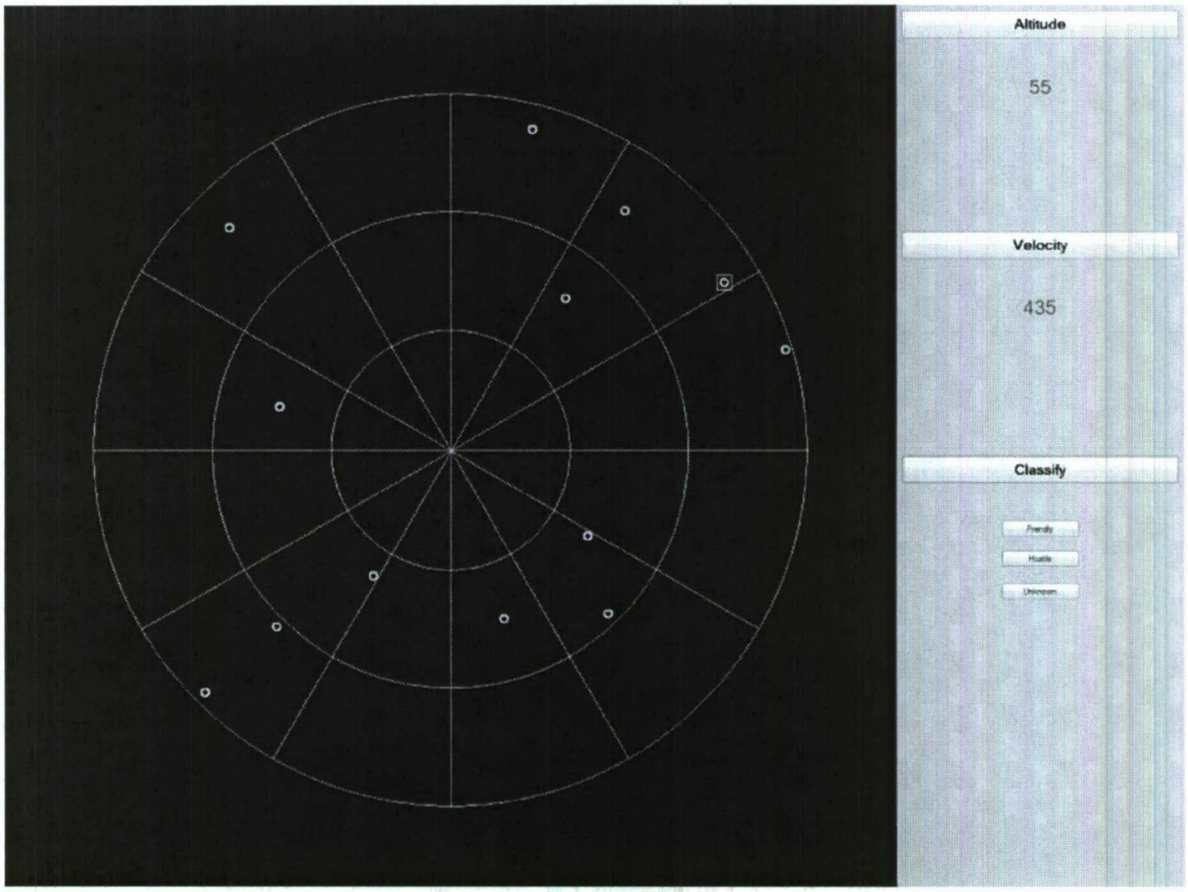
The first research phase focused on the development of a multiple goal, reinforcement learning model of our experimental platform and comparing human performance level and variability to optimal performance. Our research platform, TEAMSim (Team Event-Based Adaptive Multilevel Simulation), is a PC-based simulation of a radar-tracking task (DeShon et al., 2004; Kozlowski & DeShon, 2004). As shown in Figure 1 below, the basic task consists of a simulated radar display where contacts with different priorities and movement patterns appear that require a configurable set of decisions to process. The task may be configured to run in either team (3 person teams) or individual modes. In the individual mode, a single person operates the task in a virtual team mode where the individual is responsible for particular contact types or a section of the radar display and virtual team members are responsible for the remaining contact types or display regions. Individuals are responsible for processing targets in their own sector and for assisting teammates by processing team member contacts when the teammate is overloaded. Individuals need to "hook" contacts on the radar screen, collect information to classify their characteristics, and render an overall decision (take action or clear) for each contact. Individuals also need to learn the skills required to prevent contacts from crossing two perimeters located on the radar screen and to determine which contacts are of higher priority and should be processed first. Assisting a teammate requires the diversion of effort away from one's own sector of responsibility, necessarily incurring costs. This dynamic choice process requires the individual to either focus effort on meeting individual goals or on helping teammates to meet team goals and represents a fundamental resource allocation question that is the focus of our research.

Figure 1. TEAMSim display showing individual (yellow) and team (blue) contacts.



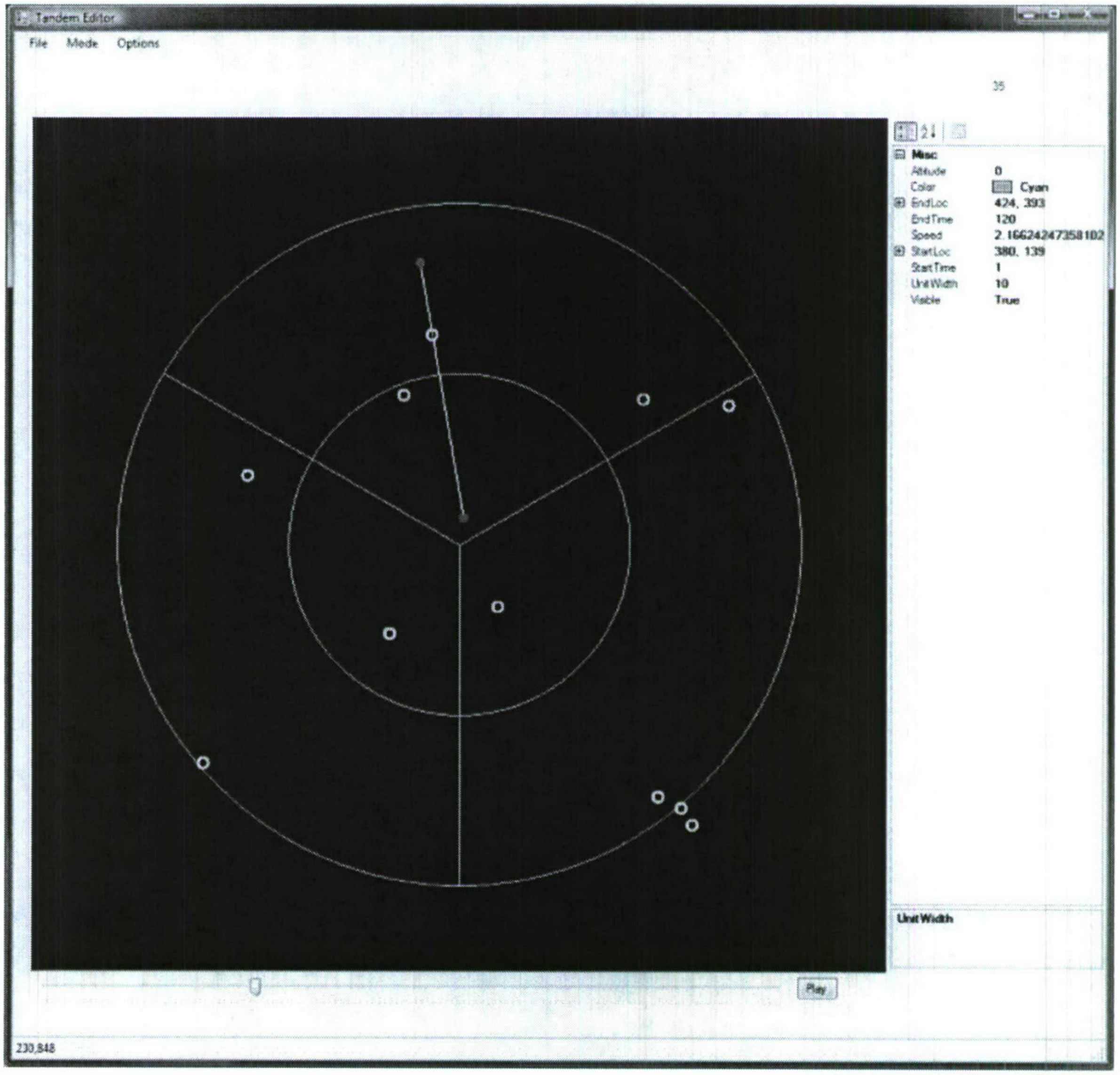
To model human learning and multiple goal regulation on our radar tracking task platform using reinforcement learning methods we developed a reduced state space version of our experimental task that is amenable to the Q-learning algorithm described above. The radar display was discretized into distinct segments as represented in Figure 2. Each segment in this display represents a possible state and the grid is scalable so the resolution can be changed easily to simulate larger state spaces with an associated increase in computational resources. Contacts may randomly enter the radar display at any segment, but once on radar display the targets move at a constant rate toward the center of the display. From any given state, the decision maker may search for contacts by moving in any of the four cardinal directions, or take action upon an encountered contact by checking a cue value associated with it, ignoring it, or removing the contact from the radar display. Rewards are accrued by correctly ignoring low priority contacts and correctly processing high priority contacts. The relative importance of individual and teammate contacts is captured by differentially weighting the rewards associated with each contact type. The software may be easily modified to output to disk real-time actions and decisions made by the human or agent along with the current reward and environment state space.

Figure 2. Example of reduced state space multiple-goal radar tracking task.



We additionally devoted substantial effort into the development of scenario generation software that enables us to easily develop repeatable task "scenarios." The task scenarios specify the number of contacts that will be processed by the human or the agent on a given trial; the duration of the trial; and the direction, altitude, velocity, and reward values of the contacts. In addition, the scenario generation software makes it possible to flexibly configure and specify additional contact cue information to make processing decisions for contacts more or less complex. A sample screen shot of this software being used in the process of developing a scenario is provided in Figure 3. Both pieces of software were developed in C++ using .NET librarys and run on 32-bit computers using the Windows operating system.

Figure 3. Example of scenario generation software for the radar tracking task.



Misc	
Altitude	0
Color	▨ Cyan
EndLoc	424, 393
EndTime	120
Speed	2.16624247358102
StartLoc	380, 139
StartTime	1
UnitWidth	10
Visible	True

In addition to the specification and development of the scenario generation and the reduced state-space experimental platform software, we were able to begin our evaluation and implementation of reinforcement learning algorithms that could be used to learn the optimal decision policies on each experimental scenario that would serve as a baseline for comparisons to human performance. As highlighted above, the goal of the first research phase was to implement standard reinforcement learning algorithms that could learn optimal action policies for our designed scenarios. We intended to use this "standard" reinforcement learning approach to also serve as a baseline for comparing our subsequent reinforcement learning approaches that would have incorporated internal drives or goals into the reinforcement learning process. The action selection models and Q-learning algorithm highlighted above were implemented in MATLAB such that an agent could learn to navigate through an environment to achieve a single goal state. The next step in the development of the reinforcement learning algorithms was to generalize the algorithms to the state space represented in our experimental scenarios and to add the larger set of actions needed to evaluate and process contacts in this state space. Unfortunately, the limited funding for this research phase precluded our ability to further these developmental goals.

What could have been accomplished had the funding continued

Simultaneous, multiple goal regulation is the hallmark of effective human behavior. Our prior research has also demonstrated the critical need for multiple goal regulation in the support of team performance. Given the wisely acknowledged importance of this issue, it is surprising to find such a dearth of theoretical and empirical research addressing the dynamics of multiple goal regulation in both humans and agents. Had six more months of support been available to support the completion of the first research phase, we would have been able to fully implement the reinforcement algorithms in our newly developed research platform. Given the Markov structure of our task, we would also have been able to compare actual human task performance with the demonstrably optimal performance of the learned policies. This would have provided unique and important insights into the strengths and limitations of real-time, multiple goal regulation of action in humans that would be generalizable far beyond our individual and team goal regulation focus.

Had we received additional funding to support the second research phase, we believe that we could have provided a strong contribution to the both the human learning and performance literatures and the reinforcement learning and action selection literatures. Current approaches to the action selection and learning problems locate reward values in features of the environment. This is inconsistent with what we know about human action selection. For example, food has different reward value when the internal state of hunger is high but far less reward value when the hunger state is low. In other words, the reward value of action-state pairings depend upon the goal satiation levels internal to the actor. We believe that modeling the internal goal states of the actor is critical to improve our understanding of the human action selection process and that this information can advance the development of reinforcement learning algorithms to support autonomous agents.

Contributions and Future Directions

The proposed research was intended to advance basic knowledge in a number of areas. From the perspective of human learning and performance, our approach would have extended current learning models to multi-criteria optimization and would have provided an approach to rapid adaptation to environmental change (e.g., spontaneous recovery from extinction) that is major limitation of virtually every existing model of human learning. From a machine learning perspective, our approach would have provided a mechanism for substantially improving the autonomy of learning agents by incorporating goals or drives with dynamic satiation levels into the agent. As a result, the reward value associated with various actions would be a function of both the environment and the current goal state space of the learning agent. Further, our approach to adaptation, via knowledge representation and memory, represents a substantial departure from existing machine learning approaches to the adaptation problem. We believe that the approach would have raised a number of questions that would have spurred further machine learning research.

Our research program is focused on modeling the human resource allocation problem that occurs when individuals function as members of a team and must simultaneously strive to achieve both individual and team goals. This project was intended to further understanding of how individuals learn to allocate resources to multiple goals (individual and team goals in our context) and the reasons why this allocation process is frequently sub-optimal. Had the work proceeded to completion, a number of steps would have been required in future investigation to generalize the findings from the proposed research to understanding the optimal allocation of team resources as multiple team members work interdependently to accomplish both individual and team goals. The first step toward this goal would be to incorporate intelligent agents, obtained via our reinforcement learning approach, into the environmental structure so that our reinforcement learning agents and our human team members would have to function in a highly interdependent dynamic environment with previously developed intelligent agents as teammates. The next step beyond that would have been to model the simultaneous learning process that occurs as teams of naive reinforcement learning agents (and humans) learn to optimally allocate resources to simultaneously optimize both individual and team goals. We believe this line of theory development fills important gaps in the optimal modeling and human performance literature. We hope to continue pushing this unique approach to the problem of dynamic multiple-goal resource allocation.

Principal Investigators

Steve W. J. Kozlowski, Ph.D. is a Professor of Organizational Psychology at Michigan State University. His major interests focus on the processes by which individuals, groups, and organizations learn and adapt their performance to novel and challenging situations. Dr. Kozlowski's current research program is focused on self-regulated learning; team training, development, and adaptation; and the role of leaders in promoting team effectiveness. The goal of this work is to generate knowledge to promote the development of adaptive individuals, teams, and organizations. Dr. Kozlowski is a Fellow of the American Psychological Association, Association for Psychological Science, International Association for Applied Psychology, and the Society for Industrial and Organizational Psychology. He is the Incoming

Editor for the *Journal of Applied Psychology* and currently serves as an Associate Editor for the outgoing Editorial Team. He has served on the Editorial Boards of the leading journals in the field including, the *Academy of Management Journal, Human Factors*, the *Journal of Applied Psychology*, and *Organizational Behavior and Human Decision Processes*. Dr. Kozlowski received his B.A. in psychology from the University of Rhode Island (1976), and his M.S. (1979) and Ph.D. (1982) degrees in organizational psychology from The Pennsylvania State University.

Richard P. DeShon, Ph.D. is a Professor of Industrial and Organizational Psychology at Michigan State University. His research interests are in the areas of motivation and self-regulation, measurement theory, cognitive measurement (intelligence, working memory, and knowledge structures), performance measurement, and selection and classification based on individual differences. He has published papers in many of the leading psychological journals such as *Psychological Bulletin, The Journal of Applied Psychology, Psychological Methods, Journal of Experimental Psychology: Learning, Memory, and Cognition, Organizational Behavior and Human Decision Making*, and *Intelligence*. He is currently on the editorial board for the *Journal of Applied Psychology* and *Psychological Methods*. He is an ad-hoc reviewer for many journals including *Organizational Behavior and Human Decision Making, Personality and Social Psychology Bulletin, Journal of Personality and Social Psychology, Psychological Bulletin*, and the *Journal of Statistical Computation and Simulation*. He is a member of the Association for Psychological Science and the Society for Industrial and Organizational Psychology. He received his B.S. in psychology from The Ohio State University (1988), and his M.A. (1991) and Ph.D. (1993) from The University of Akron.

References

Alberts, D. S., Garstka, J. J., & Frederick, P. S. (1999). Network Centric Warfare: Developing and Leveraging Information Superiority. Washington, D.C.: C4ISR Cooperative Research Program.

Atkinson, J. W, & Birch, D. (1970). The dynamics of action. New York: Wiley.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. Psychological Bulletin, 114, 80 – 99.

Daw, N. D., Niv, Y. & Dayan, P. (2006). Actions, values, policies and the basal ganglia. In E. Bezard (Ed.), Recent breakthroughs in basal ganglia research. New York: NY, Nova Science Publishers, pp. 111 - 130.

Dayan, P. (1992). The convergence of TD(λ). Machine Learning, 8, 341–362.

Dayan, P. (2002). Motivated reinforcement learning. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, pp. 11–18, Cambridge, MA, MIT Press.

DeShon, R. P., Kozlowski, S. W., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A multilevel, multiple goal model of feedback effects on the regulation of individual and team performance in training. Journal of Applied Psychology. 89, 1035-1056.

Fu, W-T. & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. Journal of Experimental Psychology: General, 135, 184-206.

Gabor, Z., Kalmar, Z., & Szepesvari, C. (1998). Multi-criteria reinforcement learning. In Proc. 15th International Conf. on Machine Learning, pp. 197–205. Morgan Kaufmann, San Francisco, CA.

Hampton, A. N., Bossaerts, P. & O'Doherty J. P. (2006). The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. Journal of Neuroscience, 26, 8360-8367.

James, J. H. & Wagner, A. R. (1980). One-trial overshadowing: Evidence of distributive processing. Journal of Experimental Psychology: Animal Behavior processes, 6, 188 – 205.

Konadaris, G. D. & Barto, A (2006). An adaptive robot motivational system. Animals to Animats 9: Proceedings of the 9th International Conference on Simulation of Adaptive Behavior, CNR, Roma, Italy.

Konadaris, G. D. & Hayes, G. M. (2005). An architecture for behavioral-based reinforcement learning. Adaptive Behavior, 13, 5 - 31.

Kozlowski, S. W. J., & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research, and principles. In E. Salas, L. R. Elliott, S. G. Schflett, & M. D. Coovert (Eds.), Scaled worlds: Development, validation, and applications. Burlington, VT: Ashgate Publishing.

Mannor, S. & Shimkin, N. (2004). A geometric approach to multi-criterion reinforcement learning. Journal of Machine Learning Research, 5, 325–360.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. Psychological Bulletin, 117, 363-386.

Montague, P. R. & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. Neuron, 36, 265–284.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioral control. Nature, 431, 760–767.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping reinforcement learning with less data and less time. Machine Learning, 13, 103–130.

Natarajan, S. & Tadepalli, P. (2005). Dynamic preferences in multi-criteria reinforcement learning. Proceedings of The 22nd International Conference on Machine Learning (ICML). Bonn, Germany.

Pearl, J. (2000). Causality. Cambridge University, New York.

Puterman, M. L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, New York.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning II: Current Research and Theory (Eds Black AH, Prokasy WF) New York: Appleton Century Crofts, pp. 64-99.

Robbins, S.J. 1990. Mechanisms underlying spontaneous recovery in autoshaping. Journal of Experimental Psycholgy: Animal Behavior Processes, 16, 235 - 249.

Schmidt, A. M. & DeShon, R. P. (in press). What to do? The effects of goal-performance discrepancies, superordinate goals, and time on dynamic goal prioritization. Journal of Applied Psychology.

Schultz, W., Dayan, P., & Montague, P. R. (1997) A neural substrate of prediction and reward. Science, 275, 1593–1599.

Schultz W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron, 36, 241-263.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317 - 1323.

Simon, H. (1956). Rational choice and the structure of the environment. Psychological Review, 63, 129-138.

Smith, A., Li, M., Becker, S. and Kapur, S. (2006), Dopamine, prediction error, and associative learning: a model-based account. Network: Computation in Neural Systems, 17, 61-84.

Sprague, N. & Ballard, D. (2003). Multiple-Goal Reinforcement Learning with Modular Sarsa(0). International Joint Conference on Artificial Intelligence (IJCAI), Acapulco, Mexico.

Spirtes, P., Glymore, C., & Schines, R. (1993). Causation, Prediction, and Search. Springer-Verlag, New York.

Sutton, R. S. & Barto, A. G. (1998). Reinforcement learning. Cambridge, MA: MIT.

Watkins, C. & Dayan, P. (1992). Q-Learning. Machine Learning, 8, 279 – 292.

Wise, R. A. (2004). Dopamine, learning and motivation. Nature Reviews Neuroscience, 5, 483 - 494.